Advanced Audio Interface for Phonetic Speech Recognition in a High Noise Environment

SBIR 99.1 TOPIC AF99-103 PHASE I SUMMARY REPORT

Prepared By

Standard Object Systems, Inc.
January 2000

for

Air Force Research Laboratory
Contract No. F41624-99-C-6019

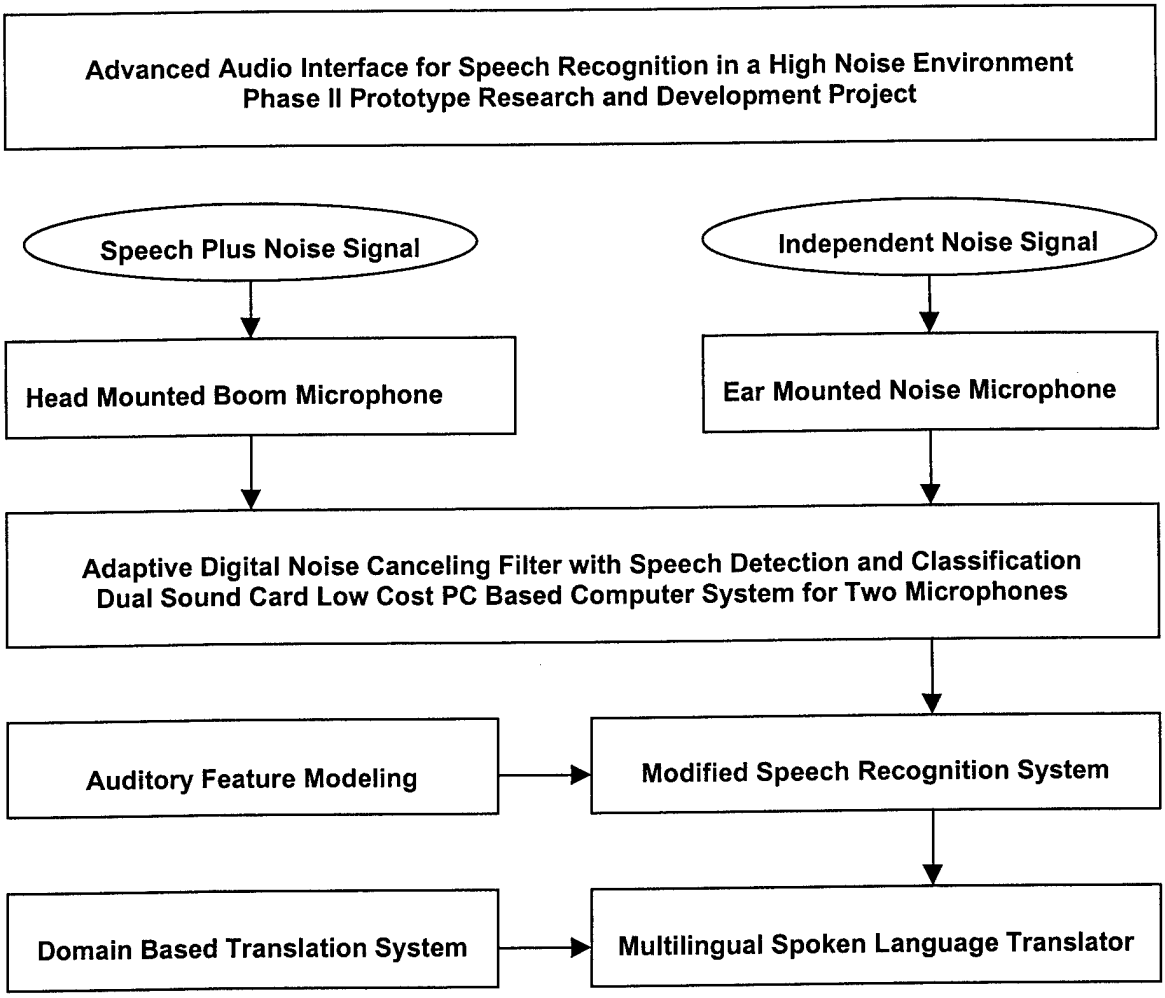
ABSTRACT

Standard Object Systems, Inc. (SOS) has used its existing technology in phonetic speech recognition, audio signal processing, and multilingual language translation to design and demonstrate an advanced audio interface for speech recognition in a high noise military environment. The Phase I result was a design for a Phase II prototype system with unique dual microphone hardware with adaptive digital filtering for noise cancellation which interfaces to speech recognition software. It uses auditory features in speech recognition training, and provides applications to multilingual spoken language translation. As a future Phase III commercial product, this system will perform real time multilingual speech recognition in noisy vehicles, offices and factories. The potential market for this technology includes any commercial speech and translation application in noisy environments.

To reduce the effect of noise in speech recognition, SOS created an adaptive digital filter to remove noise from speech prior to recognition processing. Six different filters were programmed and demonstrated using a unique dual microphone input computer system. These results were evaluated using three commercial speech recognition software systems and the SOS phonetic speech recognition (SPSR) tool kit. The maximum improvement in spoken words recognized was over 300% for the highest noise level. In addition, SOS researched computational models for noise resistant speech features using three auditory models: the auditory physiology simulation (APS), the ensemble interval histogram (EIH), and the auditory image model (AIM). These models will be implemented and tested in Phase II and will only change the speech recognition systems when they increase the existing performance. In conjunction, SOS has developed a unique multilingual speech translation method that will be implemented in Phase II for use in noisy spoken communication applications. During Phase II, a prototype unit will be developed in the first year that includes dual microphones, adaptive filtering, auditory features, speech recognition, and multilingual translation. In the second year this prototype will be tested in operational environments and improved to become a commercial product in Phase III.

Purpose of the Research Work

Both the Department of Defense and commercial users have identified the failure of computer speech recognition in noisy real-world situations as a critical problem. This Phase I SBIR research report describes the creative concept, the component designs, and the experiment analysis of a dual head-mounted noise canceling microphone with an adaptive digital filter for speech recognition in noisy environments designed by Standard Object Systems Inc. (SOS). The primary research goal is to improve computer speech recognition in noisy environments with innovative adaptive filtering of dual microphone inputs processed through dual sound cards. The following figure illustrates the sound signal processing and adaptive digital filtering designed to improve noisy speech recognition, as demonstrated in the proof of concept presentation.



SOS has applied its existing technology in phonetic speech recognition, digital signal processing, software development, audio hardware engineering, and acoustic speech science to perform this research. Numerous experiments were conducted with adaptive filters, noise models, speech recognizers, auditory feature models, and PC dual microphone hardware designs. The results of the Phase I effort are documented in the Phase I report, in the accompanying CD ROM of sound files and data, and in the proof of concept demonstration conducted at the AFRL in Wright Patterson AFB.

Research Conducted During the Phase I Project

The Phase I report provides an in-depth presentation of the design of the adaptive digital filters and the performance results from the SOS experiments for noisy speech recognition. It addresses hardware design and software interface conditions for using multiple PC microphones and low cost sound cards. The research compares several available PC speech recognition software products used to test the filtered speech signals, and presents an analysis of the test results. In addition, SOS researched three physiological models and auditory feature models of hearing for Phase II incorporation within the SPSR tool kit and other commercial speech recognition systems. The report describes the operation of the SPSR tool kit and the modifications necessary to use auditory feature data for noisy speech recognition.

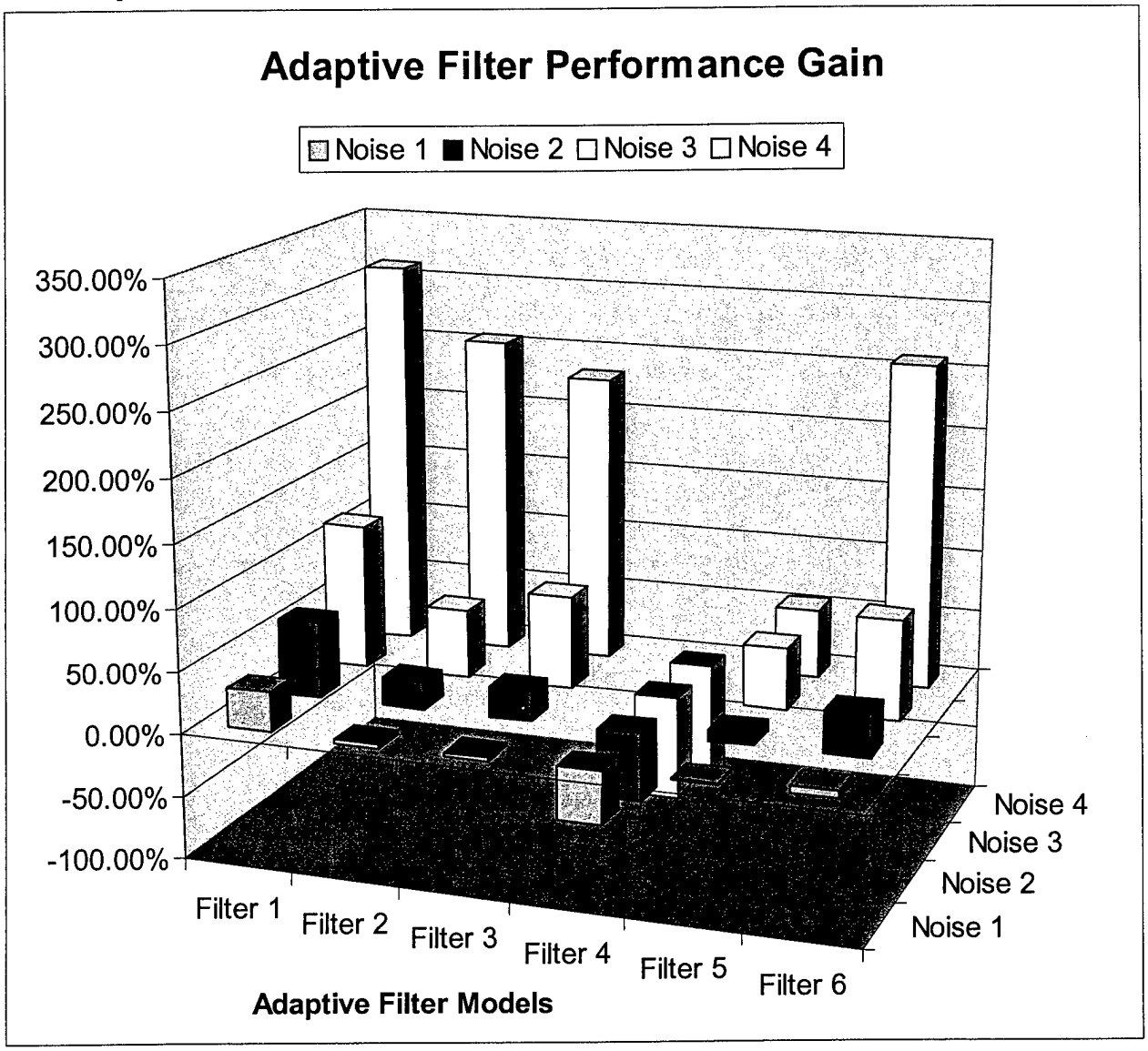
During the Phase I design, SOS constructed a number of adaptive filter experiments to prove and demonstrate the feasibility of this noise canceling technology. The following table classifies the six prototype filters by linear and nonlinear computation, time-domain and frequency-domain processing, the number of microphone inputs, and the unique speech signal reconstruction method.

NUM	NAME	L/NL	TD/FD	MIC	RECONSTRUCTION
1	Linear Adaptive Filter Bank	L	FD	2	Triangular IFFT Coherent
2	Magnitude FFT	NL	FD	2	Original Phase + Mag
3	Log Magnitude FFT	NL	FD	2	Original Phase + Mag
4	LMS ALE	L	TD	1	None, Time Shifted Output
5	Mag FFT with Iterative Recon	NL	FD	2	Phase Iteration
6	Iterative Recon with Spectral Sub	NL	FD	1	Noise Est by Scale Function

The Phase I report discusses each of these prototype adaptive filters in detail. Filter 1 is a linear adaptive filter that adjusts FIR coefficients per frequency bin for overlapping signal data blocks. In general the signal filtered with Filter 1 performs better for speech recognition programs than it sounds to the human ear. Filter 2 is a magnitude FFT experiment that accepts two sound signal inputs and produces a filtered speech signal as output. Filter 3 is a log magnitude FFT version of Filter 1 that scales the noise by the logarithm of the magnitude to approximate the response characteristic of the human ear. In general both of these filters sound clear but are not easily classified by the speech recognizer programs. Filter 4 is a least mean square adaptive line enhancer filter that uses N speech samples to predict 2 * N samples ahead. It is a linear time-domain computation that removes voiced speech from noise. The filter is limited to voiced speech signals and performs poorly with the speech recognition programs. Filter 5 is a magnitude FFT adaptive filter that scales by the noise magnitude and uses iterative reconstruction for overlapping dual signal stream data blocks. Filter 6 uses the Filter 5 computation to remove the estimated noise from a single signal input stream. In general, speech sounds reconstructed by Filters 5 and 6 are perceived more clearly by the human ear, and are easily classified by the speech recognition algorithms.

SOS investigated three separate auditory models for application to noisy speech recognition. Each uses physiological models of hearing to enhance speech recognition features, but each has a different computational basis. These contrast the statistical methods that use signal processing to transform speech into features that can be used to train and test a pattern-based recognition system. The first method is auditory physical simulation, APS, developed by SOS. This model uses continuous differential equations to represent the coupled components in the hearing process. The second method is a published ensemble interval histogram, EIH, model of the cochlea and hair cell transduction to create numerical features. The third model is the auditory image model, AIM, developed by Roy Patterson and others at Cambridge University. In each

case SOS has developed or acquired a computer program used to analyze numerical feature data in phonetic speech recognition. A major task in Phase II will be to program and analyze these models to create phonetic speech feature data. This data will be used to train speech recognition systems in the presence of noise. If replacing the existing feature models improves the speech recognition accuracy, then auditory-based features will be utilized in the adaptive noise filter design. The auditory-based features research is low risk; and will only be implemented if it enhances speech recognition accuracy, therefor auditory based features were not prototyped for the Phase I proof of concept demonstration.

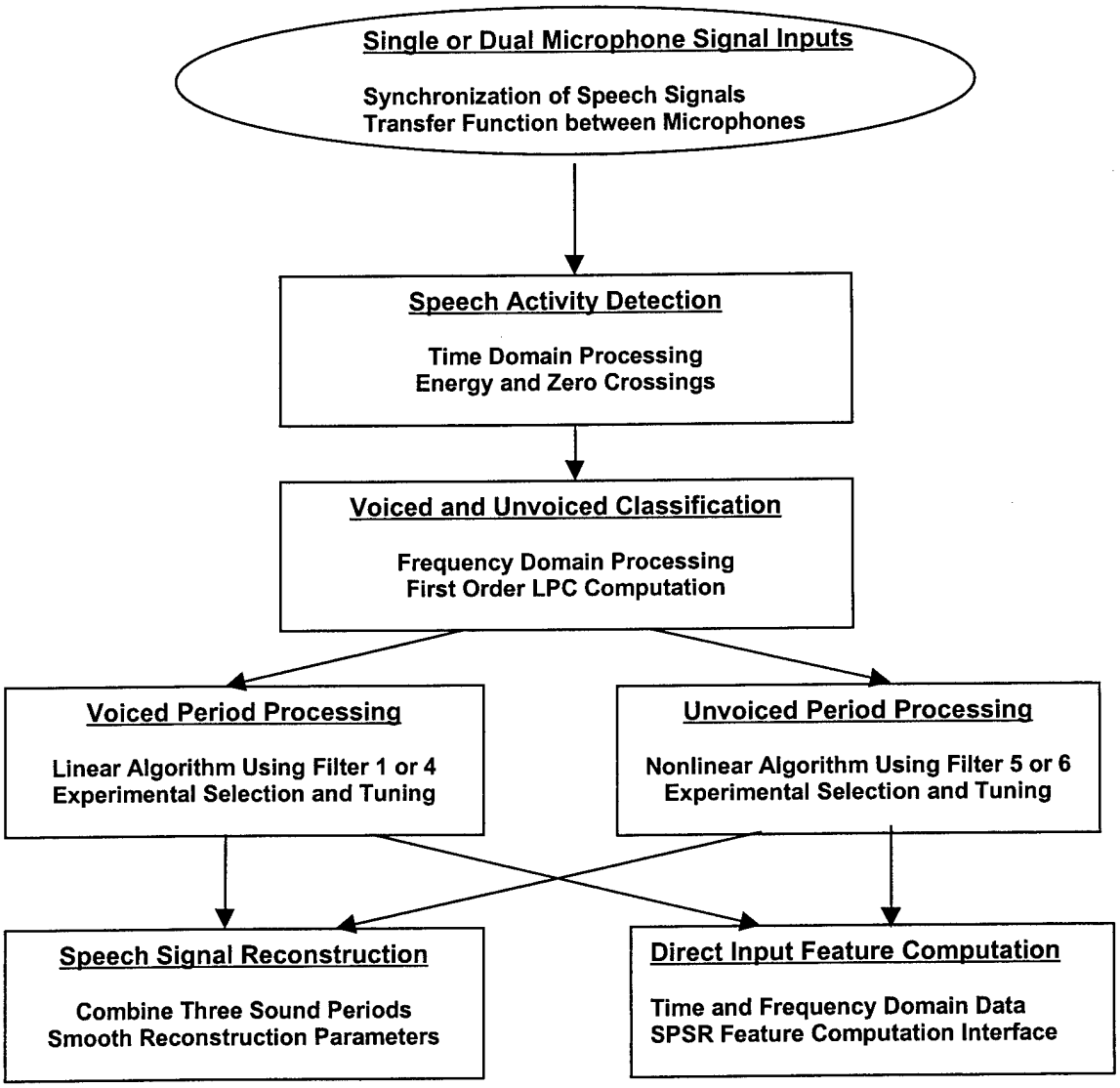


Results of the Phase I Research

The goal of the Phase I is to create a baseline design for the Phase II noise canceling filter prototype development. Each of the six prototypes was tested with the same set of speech and noise signal files to produce a sound input file for speech recognition as shown in the above figure. Four speech recognizers were selected for testing against each of the sound files to determine the correct words-recognized percentage. The figure shows the performance gain in word recognition improvement using the SRI Nuance speech recognition system for each filter

and each noise level. Other systems and other data were tested. Based on the results of the speech recognition performance and on subjective listening to the sound files, a baseline design for a Phase II adaptive noise canceling filter for speech recognition performance enhancement was created.

The implementation of the adaptive filter in Phase II will be based on the Phase I prototypes and test results. Two configurations are planned for delivery during year one of the program. First, an alpha configuration that executes on a desktop computer to provide test and evaluation data. The second configuration is a beta version, implemented for real-time execution in an operational environment for test and evaluation during year two of the project.



The design is a combination of the results from the Phase I proof of concept filter experiments with six processing components. The first two are the speech activity detector and a voiced/unvoiced state classifier. Third is a linear filter for voiced period speech processing. Forth is a nonlinear filter for unvoiced period speech processing. In the fifth, a signal reconstruction recombines the three sound segments, silence, voiced, and unvoiced into a low noise speech signal. Alternatively, in six, the sound period data can be accessed directly to

compute speech recognition feature data. This can be accomplished by a direct access to the recognition process, as in the SOS SPSR tool kit, or through specialized programming for the HTK or Sphinx II systems as an example. Most commercial speech recognition programs do not allow access to such a level of training detail. The overall filter design is similar to the front end processing used in low bandwidth sound compression systems. The combination of these six component stages will result in the development of an innovative and unique noise cancellation system targeted specifically to speech recognition enhancement.

Applications of this Research

Commercial speech recognition applications have existed for over twenty years, while computer research dates back more than forty years. Industries as numerous and diverse as banking, dictation, inspection, mail, and medical data entry use speech recognition, as shown in Figure 4-1. The most prevalent noisy environments are vehicles such as airplanes, ships, autos, tanks, and specialized industrial equipment. These environments usually require a hands-free and eyes-free speech recognition application so the operator can enter data. One example would be a policeman speaking license numbers into the digital radio system of the national crime database while driving in traffic.

Figure 4-1 Applications for Phonetic Speech Recognition with Noise

APPLICATION	REQUIREMENT	NOISE
Critical Communication (911, ATC, Police)	Intuitive operation	Line, caller environment
Routing (touch tone menus, telephones)	Speaker independent	Bandwidth limited
Control (wheelchairs, VCRs, PCs)	Reliable and accurate	Outdoor, commercial
Dictation (letters, reports, records, forms)	Large vocabulary	Office, multi speakers
Translation (languages, understanding)	Multi lingual	Office environment
Training (simulations, vocal procedures)	Complex interaction	Vehicles, machinery

Phonetic speech recognition systems, such as the SPSR Tool Kit, can be commercially applied in each of these applications. Robust speaker independent recognition is already replacing the frustrating touch tone voice menus in many telephone answering systems. As speech recognition becomes more widespread, specialized digital speech microphones will become a standard input tool alongside the mouse and keyboard. Operation in noisy environments with multiple speakers and multiple languages will be taken for granted. Figure 4-2 gives the SOS approach for the requirements identified for an advanced audio interface for noisy environment speech recognition and spoken language translation applications.

Figure 4-2 Advanced Audio Interface for Noisy Environments Requirements

REQUIREMENT	SOS APPROACH
Noisy Environments	Tested up to 90 dB live noise and speech inputs
Noise Cancellation	Adaptive digital filtering with two independent inputs
Multiple Sound Inputs	Dual microphones using two low cost PC sound cards
Speech Recognition	Utterance processing from silence period to silence period
Noise Resistant Features	Auditory based feature models and computations for training
Speaker Independent	Phonetic based with multilingual dialects, and accents
Multi Lingual Capability	Phonetic alphabet covers over 350 spoken languages
Spoken Translation	Automatic translator generation for specific domains
Interactive Visual Tools	Grammar definition, phonetic word entry, noise models
Speaker Adaptation	Session to session file to improve recognition performance
Speaker Enrollment	Simple enrollment process for speaker dependent recognition